* Use the restaurant tips file for the analytics using Excel

Download the Excel file and save it with an appropriate name for further analysis.

The file name is Restaurant tips dataset.xlsx

* Find out if there are any missing values and clean the data
  + Find the Missing Data
    - Select the data using the *ctrl+shift+right* arrow key and then the *ctrl+shift+down* arrow key. (SS1)
    - Then on the *Home tab*, in the *Editing group*, Select *Find & Select> Go to Special. (SS2)*
    - In the Go to Special dialog box, select Blanks and click OK. This will select all empty cells. (SS3)
    - As per the message “No Cells are found”. Therefore, **no missing data** is there. (SS4)
  + Data Cleaning
    - Get Rid of Extra Spaces
      * Use trim() combined with clean() to remove extra spaces, line breaks, and other unwanted characters from the whole dataset. (SS5)

TRIM(CLEAN(“cell reference”))

* + - * Replace the original values with the trimmed values using *Paste Special* > *Values. (SS6)*
      * *Remove the extra data from the sheet. Data is trimmed.*
    - Remove Duplicates
      * Select the whole data and then on the *data tab*, in the *data tools group*, Select *Remove Duplicates. (SS7)*
      * Select all the columns and the header option on the Remove Duplicates dialogue box and click on OK. (SS8)
      * There is only one duplicate row, that will be removed automatically. The remaining 243 rows are unique. (SS9)
    - Delete all formatting
      * Select the data set.
      * Go to Home –> Clear –> Clear Formats (SS10)
      * All the formatting will be removed.
    - Spell Check
      * Select the dataset.
      * Go to Review-> Proofing group-> Spelling or press F7. (SS11)
      * All the spellings will be checked. You can change or ignore it as per your requirements.
      * In the given dataset, there are two typos total\_bill (which is changed) and Thur(which is ignored). (SS12a and SS12b)
      * Then your dataset is without any spelling mistakes. (SS13)
    - Change case- Proper
      * Change the case of the dataset to proper case by using the formula

PROPER(“cell reference)

And apply it to the whole dataset. (SS14)

* + - * Replace the original values with the changed case values using *Paste Special* > *Values. (SS15)*
    - Convert numbers stored as text to numbers
      * Select the data to be converted and from the validation box choose Convert to Number. (SS16)
      * The whole data is in the correct format now. (SS17)
    - Error Checking
      * Select the dataset.
      * Press F5-> Select Special from the Go To dialogue box. (SS18)
      * From the Go To Special dialogue box. Select Formulas-> Errors. Press OK. (SS19)
      * There are no errors in the dataset. (SS20)
    - **The data is cleaned**.
* Find the features that are independent and dependent

Dependent Variable: **Tip**

Independent Variables: **Size and Total Bill**

As we can see, the tip is the dependent feature and after performing an analysis I have found that only Size and Total Bill are the variables that impact the value of the Tip.

However, I have analyzed all the columns as independent variables to justify my choice.

* Identify which predictive problem is needed.

As there are multiple independent variables and a single dependent variable. So, **Multiple Linear Regression** is best suited for this dataset.

* Encode the categorical variables to numeric values using IF conditions
  + As per the independent variables I have selected. The columns are already numeric.
  + But for converting the data to numeric the following steps are done.
    - There are 4 such columns, Sex, Smoke. Day and Time
    - First get the unique values in each categorical variable.
      * Use UNIQUE() on each column
        + UNIQUE(A2:A244)

UNIQUE(B2:B244)

UNIQUE(C2:C244)

UNIQUE(D2:D244)

* + - Then convert the values into numeric
      * Use IFS() on each categorical column
        + IFS($A2=$R$7,0,$A2=$R$8,1)
        + IFS($B2=$S$7,0,$B2=$S$8,1)
        + IFS($C2=$T$9,0,$C2=$T$10,1,$C2=$T$8,2,$C2=$T$7,3)
        + IFS($D2=$U$7,0,$D2=$U$8,1)
    - All the values are now numeric (SS21). Save the data in another sheet for further analysis.
* Build an appropriate model with the dataset.
  + Go to the Data tab-> Click on Data Analysis from the Analysis group.
  + From the Data Analysis dialogue box. Select Regression and Click OK.
  + The regression dialogue box will appear.
    - In Input Y Range, select the range of dependent variable and all the independent variables will come in Input X Range.
    - Also, Select Labels are our data has labels in the first row.
    - To get the results in different worksheet, select New Worksheet Ply.
    - Also Select Residuals, Standard Residuals and Normal Probability Plots (SS22)
  + The Output will appear on another worksheet with Summary Output, Residual Output, and Probability Output.
  + The Summary Output is divided into 3 Parts:
    - Regression Statistics: The term Adjusted R Square is of importance as the adjusted R-squared increases when the new term improves the model more than would be expected by chance. When all other variables are considered as independent this value is 0.454644643 and the values are improved and become 0.462425409 when the size and total bill are considered as independent variables.
    - ANOVA: **Significance F** is a crucial term to find the output of your model whether it is statistically significant or not. When the value of the **Significance F** is less than 0.05, the independent variables have a statistically significant relationship with the dependent variable. In both cases, this value represents there is a **significant relationship** between the independent (x values) and dependent (y values) variables.
    - Multiple Regression Coefficients: In this, **the P-value**shows the statistically significant relationship between the independent and dependent variables. Here, the **P-value** for size and total bill is below 0.05. So, only these two variables are **statistically significant** with the tips.
  + So as per all the values from the summary output, independent variables considered are: Size and Total Bill. (SS23 and SS24)
* Calculate the predicted and actual tip values.
  + As I am considering Size and Total Bill as the independent variables.
  + The equation will become:

y(Tip)=0.672163791744246+0.192346316055223\*Size+ 0.092637395155248\*Total Bill

* + The values can be predicted from this equation (SS26) and the Residual Output table also gives one column “Predicted Tip” (SS25). Both values are the same (SS27).
* Calculate the RMSE(Root Mean Square Error) of the model. RMSE is root of mean of square errors.
  + Calculate the difference between the predicted values and the actual values and create a new column Difference. (SS28)
  + Then find the Mean Square Error (MSE) using the following formula on the difference column (SS29):

SUMSQ(J2:J244)/COUNT(J2:J244)

* + The find the square root of MSE to calculate RMSE. The value of RMSE is **1.00918843042401**. (SS30)